Scene-Aware Spatiotemporal Generalization: Towards Robust Temporal Action Detection Across Domains

Fangming Feng*, Sihang Cai*, Zequn Xie, Yangyang Wu†, Tao Jin

Zhejiang University
{fangmingfeng, jint_zju}@zju.edu.cn

Abstract

Temporal Action Detection (TAD) aims to identify specific actions in long, untrimmed videos by determining their start, end times and categories, yet existing models suffer from performance degradation under out-of-distribution scenarios due to unrealistic i.i.d. assumptions. While domain generalization (DG) offers a promising solution, image-based DG methods fail to address the unique spatiotemporal challenges in video-based TAD, including the spatiotemporal complexities and significant variations in action instance scales and densities across domains. To bridge this gap, we propose the first DG framework tailored for TAD. We propose Scene-Aware Video Segmentation, which segments videos based on semantic similarity, addressing cross-domain action instance density and scale discrepancies. Additionally, we present Temporal-Aware Normalization Perturbation to generate diverse video features while preserving temporal integrity. We establish the first DG-TAD benchmark, evaluating 11 state-of-the-art DG methods across four datasets. The experiments demonstrate that our framework consistently outperforms existing approaches, achieving superior generalization on unseen domains. The proposed modules are architecture-agnostic, offering plug-and-play compatibility for broader video understanding tasks.

1 Introduction

Temporal Action Detection (TAD) is a fundamental task in long-form video understanding, aiming to identify the start/end times (t_{start}, t_{end}) and categories of actions in long untrimmed videos (Wang et al. 2024a). While deep learning has driven significant progress in TAD on specific benchmarks (Lin et al. 2024b), this success heavily relies on the unrealistic assumption that training and test data are independently and identically distributed (i.i.d.). This assumption fails in real-world applications, which are often subject to distribution shifts, posing a significant Out-of-Distribution (OOD) challenge that limits the model's generalization ability and practical value.

To tackle this challenge, Domain Generalization (DG) has emerged as a promising solution. Its objective is to train a model on source domains datasets that can generalize well to unseen target domains without accessing their data. DG has achieved remarkable success in tasks such as image classification (Su et al. 2024; Kim and Han 2024), object detection (Fan et al. 2023), and short trimmed video based action classification (Yao et al. 2022; Lin et al. 2023a).

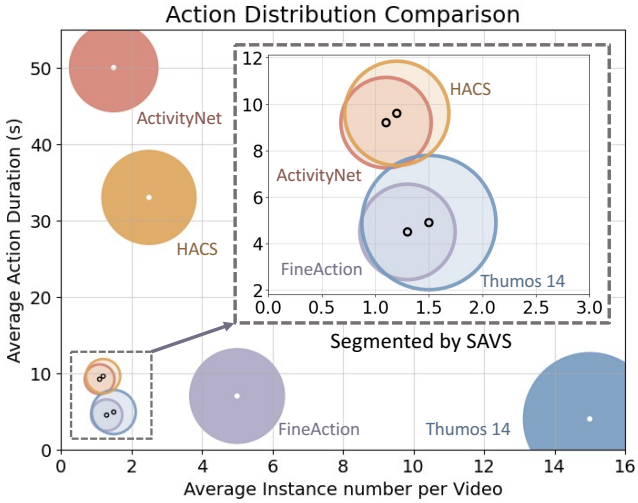

Figure 1: Comparison of action scales and densities before and after segmentation, with the x-axis representing the average instance number per video, the y-axis representing the average action duration, and bubble size indicating the average video length. The solid circles represent the original dataset, while the dataset segmented by SAVS is filled with a transparent color.

However, our investigation reveals that existing DG methods consistently underperform on TAD task (see Table 1 for details). Through an in-depth analysis, we identify a unique domain gap beyond common stylistic variations cross TAD datasets: the Action Instance Distribution gap. Specifically, TAD datasets exhibit vast discrepancies in action duration $t_{dur} = t_{end} - t_{start}$ and action density (the number of instances per video), with differences even exceeding 10-fold (Figure 1). This disparity severely impairs model generalization. For instance, a model trained on THUMOS, which features dense, short actions, tends to generate numerous short proposals when tested on ActivityNet, which contains

*These authors contributed equally.
†Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

sparse, long actions. This mismatch leads to a sharp performance drop. We argue that the failure of existing DG methods to address this critical, task-specific challenge is the primary reason for their poor performance in TAD.

Moreover, many DG methods generate diverse data via augmentation or synthesis to learn domain-invariant features. For long videos, however, this approach is problematic. Frame-by-frame augmentation is computationally infeasible. Even with efficient feature-level augmentation methods, such as perturbing channel statistics to change style (Huang and Belongie 2017; Yan et al. 2025; Fan et al. 2023), they corrupt the essential temporal dynamics in videos. This disruption, while benign in image tasks, leads to the loss of critical information (see Section 3.3 for details).

To address these challenges, we propose the first Domain Generalization framework specifically designed for TAD. Our framework tackles the core problems from two perspectives: 1) Mitigating the Action Instance Distribution Gap: We introduce a Scene-Aware Video Segmentation (SAVS) module. It intelligently segments long videos into contextually complete, single-scene clips based on a self-similarity matrix. As shown in Figure 1, SAVS effectively aligns action durations and densities across domains. 2) Generating Diverse Data while Preserving Temporal Information: We propose Temporal-Aware Normalization Perturbation (TANP). This method decouples video features into static (background) and dynamic (action) components. By applying distinct perturbations to each, TANP synthesizes style-diversified data without corrupting critical dynamic information, thereby enhancing generalization for long-video task.

To fill the research gap in DG for TAD, we establish the first DG-TAD benchmark. We systematically evaluate 11 state-of-the-art DG methods, and our results demonstrate that our proposed method significantly outperforms them across various settings. Our main contributions are summarized as follows:

- We propose the first Domain Generalization framework specifically designed for the Temporal Action Detection task, effective for both single-source and multi-source generalization.
- We introduce a Scene-Aware Video Segmentation (SAVS) module, a model-agnostic, plug-and-play unit for long-video tasks that effectively aligns cross-domain action instance distributions.
- We propose a Temporal-Aware Normalization Perturbation (TANP) method that synthesizes diverse domains without compromising temporal consistency by decoupling and selectively perturbing features.
- We establish the first DG-TAD benchmark and demonstrate the superiority of our method through extensive experiments.

2 Related Works

2.1 Temporal Action Detection

Currently, most DG algorithms focus on image-based tasks, while video-based TAD tasks face more challenging issues.(Lin et al. 2024a, 2023b) Similar to object detection tasks, TAD tasks can be broadly categorized into three main types: single-stage, two-stage, and anchor-free methods.

Single-stage TAD methods predict both the temporal boundaries and action categories of each action instance simultaneously (Wang et al. 2021; Fu et al. 2024). Two-stage TAD methods, first generate action proposals using predefined temporal anchors, and then perform boundary regression and category classification on these proposals (Zhang, Wu, and Li 2022; Chen et al. 2024). The anchor-free TAD methods, in contrast, do not rely on any predefined anchors and directly regress the action boundaries (Jin et al. 2020; Guo, Jin, and Zhao 2024).

Although existing models have achieved excellent performance on TAD tasks, their effectiveness often suffers a significant decline when applied to real-world. This is primarily due to the inevitable domain gap between training datasets and real-world environments.

2.2 Domain Generalization

Domain generalization can enhance the performance of TAD models on unseen domains. Currently, the majority of DG algorithms can be categorized into two main types: domain alignment and data augmentation.

Domain alignment-based DG methods aim to learn domain-invariant representations by reducing inter-source domain discrepancies. Common approaches include: (1) Mapping features via learnable functions to minimize feature divergence (Muandet, Balduzzi, and Schölkopf 2013); (2) Leveraging contrastive loss that groups same-class cross-domain samples as positives and different-class samples as negatives (Mahajan, Tople, and Sharma 2021); (3) Employing adversarial learning across domains to extract domain-invariant features (Deng et al. 2020). However, such methods typically require multiple labeled source domains, which poses practical challenges for video applications due to high data storage and annotation costs.

Data augmentation techniques enhance model robustness by simulating domain shifts with more diverse training data, eliminating the need for multiple source domains. These methods range from simple image transformations (Otálora et al. 2019; Fu et al. 2025) to advanced techniques like style transfer (Somavarapu, Ma, and Kira 2020; Cheng et al. 2025a) and image generation (Carlucci et al. 2019; Xie et al. 2025). However, applying these augmentations at the frame level for video is computationally expensive. While more efficient feature-level augmentation methods exist like domain-mixed instance styling (Zhou et al. 2024; Cheng et al. 2025b) and domain-agnostic normalized perturbations (Fan et al. 2023), they often overlook the critical temporal dependencies inherent in video data.

Some researchers have designed DG methods for short video action classification tasks. MIDAR (Shin et al. 2024a) learns multiple representational invariances tailored to the unique characteristics of diverse domains through multi-teacher invariance distillation. VideoDG (Yao et al. 2022) focus on learning more generalizable local temporal relation features through an architecture called the Adversarial Pyramid Network. STDN (Lin et al. 2023a) discovers potential domain-invariant cues by perceiving diverse cues from both

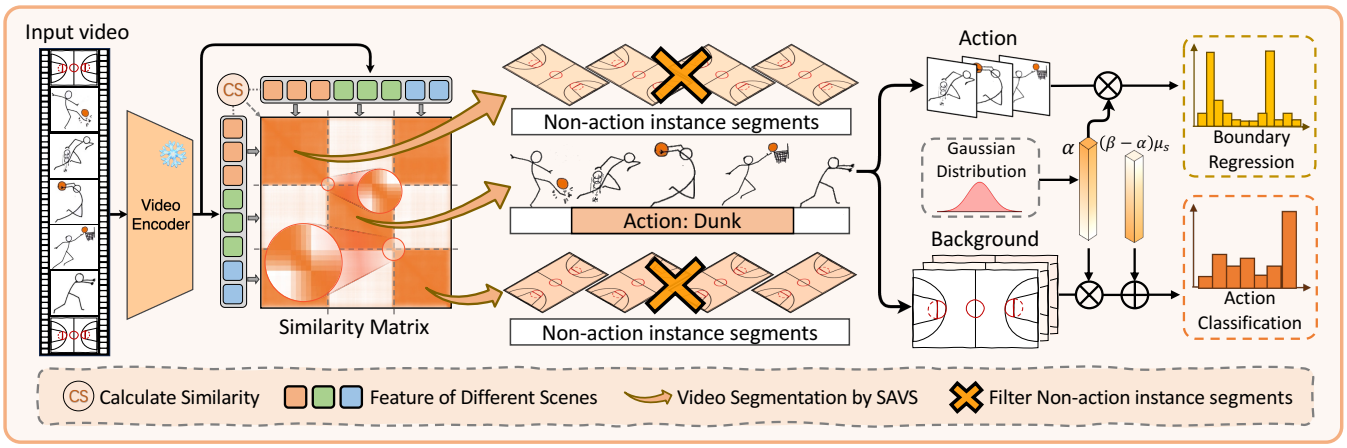

Figure 2: Overview of the proposed method. The depth of color in the similarity matrix reflects the size of the similarity values. To enhance visual clarity, we represent action videos using stick figure animations. The actual videos and their corresponding similarity matrices are provided in Appendix.

the spatial and temporal dimensions. However, these methods are only designed for short untrimmed video datasets and do not address the Action Instance Distribution gap that is unique to long untrimmed videos.

3 Method

3.1 Overview

Our framework enhances cross-domain TAD performance through three-stage representation alignment. For source domain videos, we first segment them into single scenes using semantic visual similarity matrices to normalize cross-domain action scales and density. Subsequently, non-action scenes are adaptively filtered to suppress domain-specific noise. Finally, TANP diversifies video styles through feature-level perturbations. This three-stage alignment (segmentation, filtering, perturbation) significantly reduces scale and density discrepancies of action across datasets, enhances data diversity, and consequently improves the model's domain generalization capability.

3.2 Scene-Aware Video Segmentation

We construct a feature similarity matrix $\mathbf{M} \in \mathbb{R}^{T\times T}$ to guide video segmentation, where each element $\mathbf{M}(t,t')$ quantifies the cosine similarity between features x_t and $x_{t'}$:

$$\mathbf{M}(t,t') = \frac{x_t^\top x_{t'}}{\|x_t\| \cdot \|x_{t'}\|} \in [-1,1]. \tag{1}$$

The similarity matrix exhibits a clear block-diagonal structure, with multiple square submatrices along its diagonal—each representing frames from the same scene.

Video clip features inherently encode multi-frame temporal contexts. At scene transition regions, these features blend information from adjacent scenes S_l (preceding) and S_r (succeeding) with intra-scene similarities s_l and s_r respectively. For gradual transition zone that span across n frames, we model the similarity evolution as a linear mixture:

$$s(t) = \frac{(n-t)s_l + ts_r}{n}. \tag{2}$$

For each candidate transition point $t \in [0,n]$, the analysis of feature similarities within symmetric T_{win}-sized windows reveals three distinct patterns: (1) left-dominant regions where features correlate strongly with S_l, (2) right-dominant regions with exclusive S_r correlations, and (3) transitional zones exhibiting linear blending of S_l-S_r characteristics weighted by their temporal proximity.

To ensure the analysis considers both adjacent scenes and the transition region, we define the analysis window size $T_{\text{win}} = 2(n-1)$,with $n-1$ points on both the left and right sides. The similarity distribution across T_{win} is formally expressed as:

$$s(i,t) = \begin{cases} s_l \cdot \left(1 - \frac{t}{n}\right), & (i \le 0) \\ \frac{(n-\max(i,t))\cdot s_l + \min(i,t)\cdot s_r}{n}, & (0 < i < n) \\ s_r \cdot \frac{t}{n}, & (i \ge n) \end{cases} \tag{3}$$

where i represents the relative position within the gradual transition zone. We calculate the mean of all points within the $n-1$ range around the transition point (detailed calculation process is provided in Appendix):

$$\begin{aligned} E[X] &= \frac{(n-t)(3n-t-3)\cdot \mathrm{s}_l + t(2n+t-3)\cdot \mathrm{s}_r}{4n(n-1)} \\ &= \frac{(\mathrm{s}_l+\mathrm{s}_r)t^2 + [(3-4n)\mathrm{s}_l + (2n-3)\mathrm{s}_r]t + 3\mathrm{s}_l(n^2-n)}{4n(n-1)}. \end{aligned} \tag{4}$$

The mean similarity around t forms a quadratic function, with its minimum marking the optimal transition between S_l and S_r where scene blending peaks. We analytically deter-

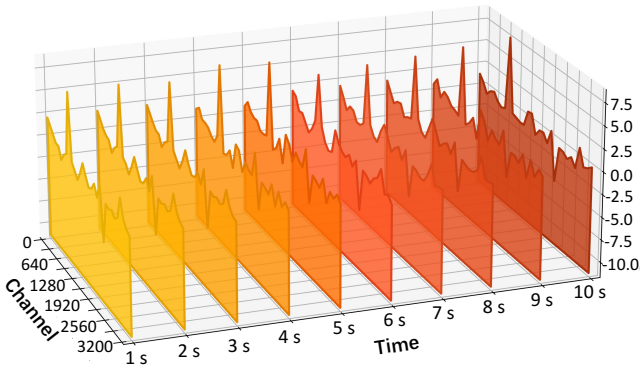

Figure 3: Visualization of video features: We extract ten features from a 10-second video and visualize them, where each feature consists of 3200 channels.

mine this point by locating its derivative's zero-crossing:

$$\frac{d(E[X])}{dt} = \frac{2(\mathrm{s}_l + \mathrm{s}_r)t + (-4n+3)\mathrm{s}_l + (2n-3)\mathrm{s}_r}{4n(n-1)}. \tag{5}$$

Setting the derivative equal to zero:

$$\frac{d(E[X])}{dt} = 0 \implies t = \frac{(4n-3)\mathrm{s}_l + (3-2n)\mathrm{s}_r}{2(\mathrm{s}_l + \mathrm{s}_r)}. \tag{6}$$

By gradual transition detection, we obtain temporally coherent video segments containing single scene content. The video segments without action instances are then filtered using the VideoMAE (Wang et al. 2023) (an action recognition model). Complete implementation specifications are provided in Appendix.

3.3 Temporal-Aware Normalization Perturbation

Analysis Prior work (Fan et al. 2023) reveals that feature channel statistics (μ_c, σ_c) correspond to latent domain styles, motivating feature-level perturbation to synthesize domains without costly video augmentation. Specifically, the feature perturbation is formulated as:

$$f' = \sigma_n \left(\frac{f - \mu_c}{\sigma_c} \right) + \mu_n, \tag{7}$$

where μ_c and σ_c denote the original channel-wise mean and standard deviation of the static scene feature. To generate plausible scene variations, the new statistical parameters μ_n and σ_n are defined as:

$$\mu_n = \beta\mu_c, \quad \sigma_n = \alpha\sigma_c, \quad \alpha, \beta \sim \mathcal{N}(1, 0.75), \tag{8}$$

where α and β are perturbation factors controlling the intensity of scene transformations.

However, unlike static data, video inherently contains temporal dependencies, and this method fails to yield significant benefits for our task. We analyze the underlying reasons for this phenomenon as follows.

The global spatiotemporal features of single-scene video clips obtained from SAVS exhibit temporal consistency throughout the video (Figure 3), capturing the unchanging scene information within the segment. However, minor temporal variations in certain feature channels indicate dynamic patterns associated with action evolution.

Video features f, extracted from a pre-trained video encoder, simultaneously incorporate background features f_b and action features f_a, such that $f = f_b + f_a$. Traditional normalization perturbation (Fan et al. 2023) computes the channel-wise mean $\mu_c = \mu_b + \mu_a$. Notably, SAVS ensures scene consistency within each segment, which implies that the background features f_b remain approximately stable across frames, meaning that $\mu_b \approx f_b$. Under this condition, the normalization perturbation of the features can be mathematically expressed as follows:

$$\begin{aligned} f' &= \sigma_n \left(\frac{f - \mu_c}{\sigma_c} \right) + \mu_n = \alpha f - \alpha\mu_c + \beta\mu_c \\ &= \alpha(f_b + f_a) - \alpha(\mu_b + \mu_a) + \beta(\mu_b + \mu_a) \\ &\approx \beta f_b + \alpha f_a + (\beta - \alpha)\mu_a, \end{aligned} \tag{9}$$

$$\Delta f = f' - f = (\beta - 1)f_b + (\alpha - 1)f_a + (\beta - \alpha)\mu_a, \tag{10}$$

here, $\alpha, \beta \in \mathbb{R}^{B\times C}$ are random noises sampled from Gaussian distribution. From Eq.(10), it can be observed that the term $(\beta - \alpha)\mu_a$ in Δf introduces an instance-dependent offset derived from the temporal average of action features. Unlike images, where the statistical values of image features encompass characteristics such as style, the temporal mean μ_a in video features reduces dynamic action patterns to a static representation. While this offset may enhance coarse-grained action recognition by emphasizing category-level information, it disrupts the temporal continuity of f_a, which is essential for precisely localizing action boundaries.

Normalization Perturbation for Video To address the limitations of indiscriminate feature perturbation, we propose a decoupled normalization perturbation framework that selectively perturbs static scene feature while preserving the temporal integrity of dynamic temporal feature.

Specifically, we begin by computing aggregated similarity scores between each feature and all others in the video sequence:

$$A_i = \sum_{j=1}^{N} \mathrm{sim}_{i,j}, \tag{11}$$

where N denotes the total number of features. The feature f_m achieving the maximum score $A_m$is identified as the main scene feature:

$$m = \arg\max_i A_i, \tag{12}$$

where m corresponds to the index of f_m, which serves as the most representative scene anchor. To disentangle static scene features, we compute the orthogonal projection of each feature f_i onto f_m:

$$f_s^{(i)} = \mathrm{proj}_{f_m}(f_i) = \left(\frac{f_i \cdot f_m}{\|f_m\|^2} \right) f_m, \tag{13}$$

where $f_s^{(i)}$ encodes the static portion of f_i aligned with main scene, representing persistent spatial context.

Next, by removing the static scene feature from all features, we obtain the dynamic temporal feature:

$$f_t^{(i)} = f_i - f_s^{(i)}. \tag{14}$$

After decoupling the features, we perturb the static scene feature f_s to diversify its channel-wise statistical distributions:

$$\begin{aligned} f_s' &= \sigma_n \left(\frac{f_s - \mu_s}{\sigma_s} \right) + \mu_n = \alpha \sigma_s \left(\frac{f_t - \mu_s}{\sigma_s} \right) + \beta \mu_s \\ &= \alpha f_s + (\beta - \alpha) \mu_s. \end{aligned} \tag{15}$$

For dynamic temporal feature perturbations, based on Eq.(10), we introduce the following constraints on the perturbation noise components to preserve the integrity of temporal information in dynamic features:

$$\alpha = \beta \implies (\beta - \alpha)\mu_a = 0. \tag{16}$$

This critical condition eliminates the temporal-disruptive mean term in the perturbation operator, enabling controlled, scaling-based perturbations of the action features, as illustrated below:

$$f_t' = \alpha_t f_t + (\beta - \alpha_t)\mu_t = \alpha_t f_t. \tag{17}$$

To ensure balanced perturbation intensity between static backgrounds and dynamic actions, we enforce parameter synchronization $\alpha_t = \alpha$. In summary, the perturbation of the video features f is specifically defined as:

$$f' = f_s' + f_t' = \alpha f + (\beta - \alpha)\mu_s. \tag{18}$$

This perturbation operator is applied after two shallow one-dimensional convolutional layers of the model.

4 Experiments

4.1 Datasets and Benchmarks

We use four widely used video datasets for TAD tasks: ActivityNet v1.3, Thumos14, FineAction, and HACS (with detailed information provided in Table 2).

These datasets were chosen for DG tasks because they cover a wide range of commonly used action categories. Due to the large scale of the datasets and the challenges in systematically summarizing their domain characteristics, we present their domain gap through feature visualization (as shown in Figure 4). Based on the significant domain gaps between the datasets, we constructed the first DG-TAD benchmark by pairing the datasets. To validate the model's generalization ability in multi-class scenarios, we also constructed domain generalization scenarios using ActivityNet v1.3 and HACS, which share 200 common categories. Given the small domain gap between them, only ActivityNet v1.3 is retained for cross-domain testing with other datasets.

4.2 Implementation Details

Our experiment conducts single domain generalization evaluation based on our benchmark, following the recommendations in (Zhou et al. 2023), using the Transformer architecture ActionFormer and the Mamba architecture ActionMamba as model frameworks to eliminate the influence of specific architectures on the method's effectiveness.

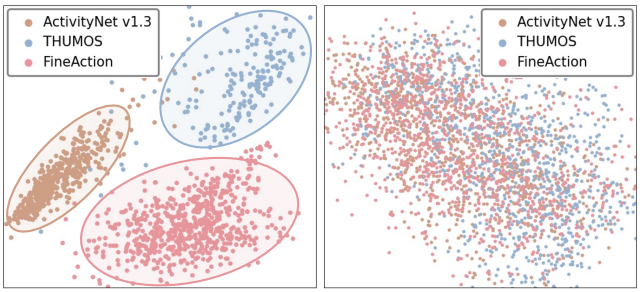

Figure 4: The visualization of the feature on ActivityNet v1.3, Thumos 14 and FineAction. The left side presents the original feature distribution, whereas the right side displays the feature distribution after SAVS and TANP.

In the experiment, all videos are standardized to 25 FPS, and during the feature extraction phase, I3D (Carreira and Zisserman 2017), InterVideo2-6B (Wang et al. 2024b) and VideoMAEv2-g (Wang et al. 2023) are used as video encoders, with snippet size and stride both set to 16. The training phase uses the Adam optimizer with an initial learning rate of 1×10^{-4}, applying cosine learning rate decay. The batch size is set to 16, with weight decay set to 1×10^{-4}. Consistent with other TAD methods (Zhang, Wu, and Li 2022; Chen et al. 2024), mAP is used as the evaluation metric, and results are reported by averaging mAP scores across temporal Intersection over Union thresholds ranging from 0.3 to 0.7 with a step size of 0.1.

4.3 Experimental Results and Analysis

To objectively assess the effectiveness of our method, we compare it with several advanced DG methods, including the basic baseline ERM, domain adversarial methods (ASFOD (Chu et al. 2023), ABA (Cheng, Gokhale, and Yang 2023)), sharpness-aware optimization series (SAM (Foret et al. 2021), UDIM (Shin et al. 2024b), LTDG (Su et al. 2024)), video task DG methods (VideoDG (Yao et al. 2022), MIDAR (Shin et al. 2024a), STDN (Lin et al. 2023a)), and other advanced methods (NP (Fan et al. 2023), RASP (Kim and Han 2024)). As shown in Table 1, our method outperforms the existing methods across all benchmark tests, especially achieving a significant advantage in the A-T and T-A groups. This is due to the significant annotation differences between the ActivityNet v1.3 and Thumos 14: the former contains on average only 1.5 action instances per video (with a duration of about 50 seconds), while the latter contains up to 15 instances (with an average of 4 seconds per instance). Traditional DG methods designed for images or short videos perform poorly because they fail to adapt to the large differences in action density and scale. We divides multi-scene long videos into single-scene short videos by SAVS, standardizing action density and scale to alleviate domain gaps, thereby significantly improving detection performance. As shown in Figure 1 and Table 2, SAVS unifies the action instance density of ActivityNet v1.3 and Thumos 14 with the largest initial density disparity—into same levels. Furthermore, the action duration gap, which originally exceeded

Methods	Years	A-T	T-A	A-F	F-A	T-F	F-T	A-H	H-A
ERM	-	$18.9_{\pm 0.4}$	$23.3_{\pm 0.4}$	$33.6_{\pm 0.2}$	$30.3_{\pm 0.4}$	$31.0_{\pm 0.1}$	$47.0_{\pm 0.1}$	$60.4_{\pm 0.2}$	$63.4_{\pm 0.2}$
ASFOD (Chu et al. 2023)	23	$18.9_{\pm 0.4}$	$25.5_{\pm 0.4}$	$35.2_{\pm 0.1}$	$33.2_{\pm 0.2}$	$33.0_{\pm 0.2}$	$50.1_{\pm 0.3}$	$60.2_{\pm 0.1}$	$62.9_{\pm 0.3}$
ABA (Cheng, Gokhale, and Yang 2023)	23	$21.8_{\pm 0.2}$	$26.5_{\pm 0.4}$	$33.4_{\pm 0.5}$	$32.7_{\pm 0.3}$	$34.4_{\pm 0.1}$	$48.9_{\pm 0.5}$	$61.4_{\pm 0.2}$	$65.9_{\pm 0.1}$
SAM (Foret et al. 2021)	21	$19.3_{\pm 0.1}$	$25.6_{\pm 0.2}$	$34.9_{\pm 0.3}$	$32.3_{\pm 0.3}$	$33.3_{\pm 0.3}$	$47.4_{\pm 0.2}$	$60.2_{\pm 0.1}$	$63.1_{\pm 0.0}$
LTDG (Su et al. 2024)	24	$21.4_{\pm 0.1}$	$24.3_{\pm 0.2}$	$33.5_{\pm 0.2}$	$29.4_{\pm 0.3}$	$33.6_{\pm 0.1}$	$46.9_{\pm 0.1}$	$59.5_{\pm 0.1}$	$65.0_{\pm 0.1}$
UDIM (Shin et al. 2024b)	24	$18.7_{\pm 0.2}$	$25.0_{\pm 0.3}$	$33.8_{\pm 0.1}$	$29.0_{\pm 0.3}$	$33.0_{\pm 0.2}$	$48.3_{\pm 0.3}$	$62.4_{\pm 0.2}$	$62.8_{\pm 0.2}$
VideoDG (Yao et al. 2022)	22	$18.8_{\pm 0.3}$	$25.6_{\pm 0.3}$	$35.8_{\pm 0.1}$	$30.5_{\pm 0.2}$	$34.8_{\pm 0.3}$	$48.4_{\pm 0.3}$	$61.0_{\pm 0.0}$	$65.6_{\pm 0.3}$
STDN (Lin et al. 2023a)	23	$20.1_{\pm 0.2}$	$25.5_{\pm 0.2}$	$35.0_{\pm 0.0}$	$31.9_{\pm 0.3}$	$34.2_{\pm 0.3}$	$47.8_{\pm 0.3}$	$62.2_{\pm 0.2}$	$60.0_{\pm 0.3}$
MIDAR (Shin et al. 2024a)	24	$19.8_{\pm 0.3}$	$25.7_{\pm 0.2}$	$34.4_{\pm 0.0}$	$32.0_{\pm 0.1}$	$33.1_{\pm 0.1}$	$49.6_{\pm 0.1}$	$62.0_{\pm 0.2}$	$63.4_{\pm 0.1}$
NP (Fan et al. 2023)	23	$20.0_{\pm 0.3}$	$26.0_{\pm 0.3}$	$35.3_{\pm 0.4}$	$32.8_{\pm 0.2}$	$33.2_{\pm 0.3}$	$49.4_{\pm 0.4}$	$61.2_{\pm 0.3}$	$64.2_{\pm 0.3}$
RASP (Kim and Han 2024)	24	$18.5_{\pm 0.3}$	$25.7_{\pm 0.1}$	$36.4_{\pm 0.2}$	$32.1_{\pm 0.3}$	$33.8_{\pm 0.1}$	$49.0_{\pm 0.3}$	$60.7_{\pm 0.2}$	$65.8_{\pm 0.3}$
Ours	-	$\mathbf{25.3}_{\pm 0.2}$	$\mathbf{30.8}_{\pm 0.2}$	$\mathbf{37.5}_{\pm 0.0}$	$\mathbf{34.9}_{\pm 0.2}$	$\mathbf{36.9}_{\pm 0.1}$	$\mathbf{51.9}_{\pm 0.2}$	$\mathbf{64.4}_{\pm 0.2}$	$\mathbf{67.6}_{\pm 0.2}$

Table 1: Comparisons with Other DG Methods. "A" denotes the ActivityNet v1.3 dataset, "T" denotes the Thumos dataset, "F" denotes the FineAction dataset, and "H" denotes the HACS dataset. We adopt ActionFormer as our TAD backbone and InterVideo2-6B as the feature encoder.

Dataset	**A**	**F**	**T**	**H**
Number of Videos	13,800	16,732	413	49,581
Number of Categories	200	106	20	200
Before segmentation				
Average Video Length	140s	150s	210s	149s
Instance number per Video	1.5	5	15	2.5
Average Action Duration	50s	7s	4s	33s
After segmentation				
Average Video Length	47.1s	49.7s	69.9s	54.3s
Instance number per Video	1.1	1.3	1.5	1.2
Average Action Duration	9.2s	4.5s	4.9s	9.6s

Table 2: Summary of Datasets Used in Experiments.

Arch	Methods	A-T	T-A	A-F	F-A	T-F	F-T
Transformer (Vaswani et al. 2017)	ERM	18.9	23.3	33.6	30.3	31.0	47.0
	ABA	21.8	26.5	33.4	32.7	34.4	48.9
	UDIM	18.7	25.0	33.8	29.0	33.0	48.3
	STDN	20.1	25.5	35.0	31.9	34.2	47.8
	NP	20.0	26.0	35.3	32.8	33.2	49.4
	Ours	**25.3**	**30.8**	**37.5**	**34.9**	**36.9**	**51.9**
Mamba (Gu and Dao 2023)	ERM	19.4	25.0	34.1	32.0	32.2	48.9
	ABA	19.5	26.9	35.2	33.3	34.0	49.6
	UDIM	19.6	27.4	35.3	32.1	34.9	49.3
	STDN	22.1	27.7	36.6	32.2	35.1	49.1
	NP	20.2	27.2	36.1	33.6	34.5	49.4
	Ours	**26.4**	**31.0**	**38.8**	**34.4**	**37.6**	**50.3**

Table 3: Comparison of the influence of Transformer and Mamba architectures on our DG algorithm.

10-fold, is reduced to less than two-fold after segmentation.

In line with the suggestions from (Zhou et al. 2023), we objectively assess the method's performance by testing various architecture models (results are presented in Table 3). Our method outperforms the comparison methods under both the Transformer and Mamba architectures. Overall, the Transformer performs slightly worse than the Mamba architecture. The former's self-attention mechanism experiences quadratic growth in computational complexity with respect to sequence length, forcing the use of local attention mechanisms for long videos, which leads to a loss of long-range dependency modeling ability. Furthermore, the size selection of the local attention window are closely tied to the data distribution, which further weakening generalization ability. In comparison, Mamba reduces computational complexity by introducing the HIPPO matrix and enhances generalization capability through a controllable memory range, resulting in superior performance in this task. Notably, by segment long videos into short ones while reducing the disparity in action duration distributions, our method reduces the computational burden of the Transformer and alleviates the sensitivity of the local attention mechanism to window size (as demonstrated in Section 4.5). As a result, even when using the Transformer architecture, its performance still is comparable to Mamba-based models.

To eliminate the potential interference of video feature extractors on the results, we validates the method using three general models from the field of video understanding. As shown in Table 4, our method consistently outperforms all feature extractors and tasks.

4.4 Ablation Study

To validate the effectiveness of the proposed components, we performs ablation experiments in the benchmark tests to explore the effects of SAVS and TANP. First, the two modules are removed individually for testing (results are shown in Table 5). The results show that SAVS significantly contributes to improving domain generalization performance, while the effect of TANP alone is limited. However, when combining SAVS to divide multi-scene videos into single-scene videos, adding TANP can significantly boost the model's generalization capability. Next, we will further analyze the effectiveness of our segmentation and perturbation methods as follows.

Impact of segmentation methods To emphasize the advantages of SAVS, we further compare it with commonly

Encoder	Methods	A-T	T-A	A-F	F-A	T-F	F-T
I3D (Carreira and Zisserman 2017)	ERM	15.2	19.8	26.1	25.2	26.0	40.8
	ABA	18.7	22.7	28.4	27.8	28.5	42.7
	UDIM	17.4	21.0	28.2	28.4	29.4	44.9
	STDN	16.9	21.8	29.8	27.2	25.0	41.5
	NP	17.8	24.3	28.5	28.5	29.2	41.2
	Ours	**21.7**	**26.7**	**30.4**	**29.0**	**30.2**	**46.0**
VideoMAE v2 (Wang et al. 2023)	ERM	18.8	22.4	32.1	30.7	29.6	47.4
	ABA	21.4	26.1	32.6	31.4	34.1	47.7
	UDIM	20.7	26.3	32.4	31.5	32.9	47.6
	STDN	20.9	27.4	33.5	32.4	32.5	46.4
	NP	21.0	26.8	35.4	33.1	32.4	48.7
	Ours	**25.1**	**30.1**	**37.4**	**35.8**	**34.9**	**50.7**
InterVideo 2 (Wang et al. 2024b)	ERM	18.9	23.3	33.6	30.3	31.0	47.0
	ABA	21.8	26.5	33.4	32.7	34.4	48.9
	UDIM	18.7	25.0	33.8	29.0	33.0	48.3
	STDN	20.1	25.5	35.0	31.9	34.2	47.8
	NP	20.0	26.0	35.3	32.8	33.2	49.4
	Ours	**25.3**	**30.8**	**37.5**	**34.9**	**36.9**	**51.9**

Table 4: Comparison of the impact of different video encoders on our methods.

SAVS	TANP	A-T	T-A	A-F	F-A	T-F	F-T
✗	✗	18.9	23.3	33.6	30.3	31.0	47.0
✓	✗	22.3	27.9	35.2	33.7	36.0	50.2
✗	✓	19.3	23.6	34.5	30.4	31.9	47.5
✓	✓	**25.3**	**30.8**	**37.5**	**34.9**	**36.9**	**51.9**

Table 5: Ablation Study on Different Components of Our Method.

used video resize methods (random segmentation, downsampling): random segmentation divides the video into 30-60 second segments, while downsampling scales the video to a uniform 60 seconds. The experimental results, shown in Table 6a, indicate that random segmentation fragments the action instances, damaging their semantics, and downsampling causes information loss when handling long videos with short action instances (such as in the Thumos 14). Our method, by using scene information, retains the integrity of action instances while effectively mitigating the issue of action distribution differences, achieving optimal performance in multiple benchmark tests.

Impact of perturbation methods To comprehensively evaluate the effectiveness of TANP, we compares not only with models that do not use normalization algorithms but also with basic normalization perturbation methods (results are shown in Table 6b). While regular normalization perturbation proves effective in most benchmark tests, it significantly impacts performance due to the disruption of video temporal information, causing a noticeable decline in the T-F benchmark test. TANP uses a feature decoupling strategy to separate the static background features and dynamic temporal action features of the video, achieving efficient and stable normalization perturbation, resulting in better and more stable performance improvement.

4.5 Parameter Sensitivity Analysis

As mentioned in Section 4.3, some hyperparameters of the Transformer architecture depend on specific datasets, affecting the model's generalization ability. Thus, we performs a parameter sensitivity analysis for the maximum sequence length and local attention window size, comparing it with the ERM method without SAVS (where long videos are downsampled when their length exceeds the maximum sequence length). The experimental results (as shown in Figure 5) demonstrate that by standardizing the video scale and action instance density, our method significantly reduces the model's sensitivity to Transformer hyperparameters and effectively enhances its generalization ability.

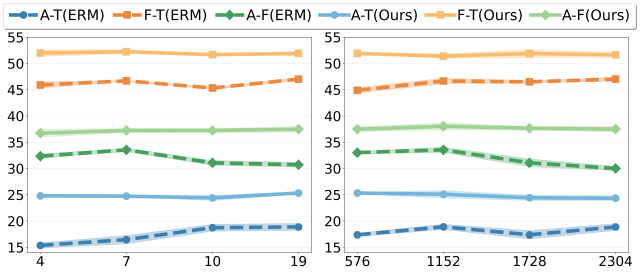

Figure 5: Sensitivity Analysis of Transformer Hyperparameters. The dashed lines represent results from the baseline ERM, while solid lines denote our proposed method. Left: Sensitivity to window size of the local attention mechanism. Right: Sensitivity to the predefined maximum sequence length of the Transformer.

w/o Split	Random	Sample	SAVS	A-T	T-A	A-F	F-A	T-F	F-T
✓	✗	✗	✗	19.3	23.6	34.5	30.4	31.9	47.5
✗	✓	✗	✗	23.1	28.5	36.3	29.7	34.1	48.7
✗	✗	✓	✗	18.9	22.4	34.9	29.4	32.6	47.8
✗	✗	✗	✓	**25.3**	**30.8**	**37.5**	**34.9**	**36.9**	**51.9**

(a) Impact of Different Segmentation Methods

w/o NP	NP	TANP	A-T	T-A	A-F	F-A	T-F	F-T
✓	✗	✗	22.3	27.9	35.2	33.7	36.0	50.2
✗	✓	✗	23.2	27.5	35.0	33.1	34.9	50.5
✗	✗	✓	**25.3**	**30.8**	**37.5**	**34.9**	**36.9**	**51.9**

(b) Comparison of Traditional NP vs. TANP

Table 6: Ablation Studies on Segmentation Methods and Perturbation Strategies

5 Conclusion

We present the first domain generalization framework for temporal action detection, addressing spatiotemporal challenges in untrimmed videos through two synergistic innovations: SAVS that aligns cross-domain action densities and scales via semantic scene segmentation, coupled with TANP enabling style diversification while preserving motion dynamics. Our method has demonstrated outstanding performance across comparative experiments, ablation studies, and parameter sensitivity experiments. These architecture-agnostic modules provide plug-and-play adaptability across long video tasks.

Acknowledgments

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang under (Grant No. 2025C02110), Public Welfare Research Program of Ningbo under (Grant No. 2024S062), and Yongjiang Talent Project of Ningbo under (Grant No. 2024A-161-G).

References

Carlucci, F. M.; Russo, P.; Tommasi, T.; and Caputo, B. 2019. Hallucinating Agnostic Images to Generalize Across Domains. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, 3227–3234. IEEE.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 4724–4733. IEEE Computer Society.

Chen, G.; Huang, Y.; Xu, J.; Pei, B.; Chen, Z.; Li, Z.; Wang, J.; Li, K.; Lu, T.; and Wang, L. 2024. Video Mamba Suite: State Space Model as a Versatile Alternative for Video Understanding. *CoRR*, abs/2403.09626.

Cheng, S.; Gokhale, T.; and Yang, Y. 2023. Adversarial Bayesian Augmentation for Single-Source Domain Generalization. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 11366–11376. IEEE.

Cheng, X.; Fu, D.; Yang, X.; Fang, M.; Hu, R.; Lu, J.; Jionghao, B.; Wang, Z.; Ji, S.; Huang, R.; Li, L.; Chen, Y.; Jin, T.; and Zhao, Z. 2025a. OmniChat: Enhancing Spoken Dialogue Systems with Scalable Synthetic Data for Diverse Scenarios. arXiv:2501.01384.

Cheng, X.; Hu, R.; Yang, X.; Lu, J.; Fu, D.; Wang, Z.; Ji, S.; Huang, R.; Zhang, B.; Jin, T.; and Zhao, Z. 2025b. VoxDialogue: Can Spoken Dialogue Systems Understand Information Beyond Words? In *The Thirteenth International Conference on Learning Representations*.

Chu, Q.; Li, S.; Chen, G.; Li, K.; and Li, X. 2023. Adversarial Alignment for Source Free Object Detection. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 452–460. AAAI Press.

Deng, Z.; Ding, F.; Dwork, C.; Hong, R.; Parmigiani, G.; Patil, P.; and Sur, P. 2020. Representation via Representations: Domain Generalization via Adversarially Learned Invariant Representations. *CoRR*, abs/2006.11478.

Fan, Q.; Segù, M.; Tai, Y.; Yu, F.; Tang, C.; Schiele, B.; and Dai, D. 2023. Towards Robust Object Detection Invariant to Real-World Domain Shifts. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Fu, D.; Cheng, X.; Li, L.; Yang, X.; Yang, L.; and Jin, T. 2025. PACHAT: Persona-Aware Speech Assistant for Multiparty Dialogue. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 29313–29330. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Fu, D.; Cheng, X.; Yang, X.; Hanting, W.; Zhao, Z.; and Jin, T. 2024. Boosting Speech Recognition Robustness to Modality-Distortion with Contrast-Augmented Prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 3838–3847. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *CoRR*, abs/2312.00752.

Guo, Z.; Jin, T.; and Zhao, Z. 2024. Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 1726–1736. Association for Computational Linguistics.

Huang, X.; and Belongie, S. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 1510–1519.

Jin, T.; Huang, S.; Chen, M.; Li, Y.; and Zhang, Z. 2020. SBAT: Video Captioning with Sparse Boundary-Aware Transformer. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 630–636. ijcai.org.

Kim, T.; and Han, B. 2024. Randomized Adversarial Style Perturbations for Domain Generalization. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, 2306–2314. IEEE.

Lin, K.; Du, J.; Gao, Y.; Zhou, J.; and Zheng, W. 2023a. Diversifying Spatial-Temporal Perception for Video Domain Generalization. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lin, W.; Chen, J.; Shi, J.; Guo, Z.; Zhu, Y.; Wang, Z.; Jin, T.; Zhao, Z.; Wu, F.; Yan, S.; and Zhang, H. 2024a. Action Imitation in Common Action Space for Customized Action Image Synthesis. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems*

2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.

Lin, W.; Feng, Y.; Han, W.; Jin, T.; Zhao, Z.; Wu, F.; Yao, C.; and Chen, J. 2024b. E^3: Exploring Embodied Emotion Through A Large-Scale Egocentric Video Dataset. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Lin, W.; Jin, T.; Wang, Y.; Pan, W.; Li, L.; Cheng, X.; and Zhao, Z. 2023b. Exploring Group Video Captioning with Efficient Relational Approximation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15235–15244.

Mahajan, D.; Tople, S.; and Sharma, A. 2021. Domain Generalization using Causal Matching. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 7313–7324. PMLR.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain Generalization via Invariant Feature Representation. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, 10–18. JMLR.org.

Otálora, S.; Atzori, M.; Andrearczyk, V.; Khan, A.; and Müller, H. 2019. Staining Invariant Features for Improving Generalization of Deep Convolutional Neural Networks in Computational Pathology. *Frontiers in Bioengineering and Biotechnology*, 7.

Shin, J.; Maiti, A.; Zou, Y.; and Choi, J. 2024a. Multi-teacher Invariance Distillation for Domain-Generalized Action Recognition. In Antonacopoulos, A.; Chaudhuri, S.; Chellappa, R.; Liu, C.; Bhattacharya, S.; and Pal, U., eds., *Pattern Recognition - 27th International Conference, ICPR 2024, Kolkata, India, December 1-5, 2024, Proceedings, Part XXIX*, volume 15329 of *Lecture Notes in Computer Science*, 116–132. Springer.

Shin, S.; Bae, H.; Na, B.; Kim, Y.; and Moon, I. 2024b. Unknown Domain Inconsistency Minimization for Domain Generalization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Somavarapu, N.; Ma, C.; and Kira, Z. 2020. Frustratingly Simple Domain Generalization via Image Stylization. *CoRR*, abs/2006.11207.

Su, H.; Luo, W.; Liu, D.; Wang, M.; Tang, J.; Chen, J.; Wang, C.; and Chen, Z. 2024. Sharpness-Aware Model-Agnostic Long-Tailed Domain Generalization. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 15091–15099. AAAI Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wang, B.; Zhao, Y.; Yang, L.; Long, T.; and Li, X. 2024a. Temporal Action Localization in the Deep Learning Era: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(4): 2171–2190.

Wang, C.; Cai, H.; Zou, Y.; and Xiong, Y. 2021. RGB Stream Is Enough for Temporal Action Detection. *CoRR*, abs/2107.04362.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 14549–14560. IEEE.

Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Chen, G.; Pei, B.; Zheng, R.; Wang, Z.; Shi, Y.; Jiang, T.; Li, S.; Xu, J.; Zhang, H.; Huang, Y.; Qiao, Y.; Wang, Y.; and Wang, L. 2024b. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, volume 15143 of *Lecture Notes in Computer Science*, 396–416. Springer.

Xie, Z.; Wang, C.; Wang, Y.; Cai, S.; Wang, S.; and Jin, T. 2025. Chat-Driven Text Generation and Interaction for Person Retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 5259–5270.

Yan, W.; Lin, W.; Guo, Z.; Wang, Y.; Feng, F.; Yang, X.; Wang, Z.; and Jin, T. 2025. Diff-prompt: Diffusion-driven prompt generator with mask supervision. *arXiv preprint arXiv:2504.21423*.

Yao, Z.; Wang, Y.; Wang, J.; Yu, P. S.; and Long, M. 2022. VideoDG: Generalizing Temporal Relations in Videos to Novel Domains. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11): 7989–8004.

Zhang, C.; Wu, J.; and Li, Y. 2022. ActionFormer: Localizing Moments of Actions with Transformers. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13664 of *Lecture Notes in Computer Science*, 492–510. Springer.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2023. Domain Generalization: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4): 4396–4415.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2024. MixStyle Neural Networks for Domain Generalization and Adaptation. *Int. J. Comput. Vis.*, 132(3): 822–836.